Tagger

Eder, Elisabeth; elisabeth.eder@aau.at

Tagger sind Programme, die Text, meistens in tokenisierter Form (*Tokenizer*), automatisch mit entsprechenden Tags aus festgelegten Tagsets annotieren. Sie basieren großteils auf *Machine Learning* und wurden auf ausgewählten Korpora trainiert, die bereits Annotationen nach bestimmten Tagsets enthalten. In vielen Fällen lassen sich die Tagger auch auf eigenen annotierten Daten trainieren, zum Beispiel auf einer neuen Sprache oder mit einem alternativen Tagset. In Bezug auf *Part-of-Speech-Tagging* sind hier der TreeTagger (Schmid 1994; Schmid 1995) sowie der neuere RNNTagger (Schmid 2019), die beide zudem die jeweiligen Lemmata der einzelnen Token ausgeben (Lemmatisierung), zu erwähnen. Neben einer Auswahl von PoS-Taggern ist auch der TreeTagger in *WebLicht* inkludiert. Der SoMeWeTa (*Social Media and Web Tagger*) (Proisl 2018) eignet sich speziell für deutsche Texte aus dem Social-Media- und Web-Bereich. Die *Python-Libraries spaCy*, *Natural Language Toolkit* (NLTK) und *flair* bieten ebenfalls PoS-Tagging an.

Literatur:

- Akbik, Alan; Blythe, Duncan; Vollgraf, Roland: Contextual String Embeddings for Sequence Labeling. In: Proceedings of the 27th International Conference on Computational Linguistics COLING. Santa Fe, New Mexico, USA: 2018, S. 1638–1649.
- Proisl, Thomas: SoMeWeTa: A Part-of-Speech Tagger for German Social Media and Web Texts. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) LREC. Miyazaki, Japan: 2018.
- Schmid, Helmut: Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing. Manchester, United Kingdom: 1994.
- Schmid, Helmut: Improvements in Part-of-Speech Tagging with an Application to German. In: Proceedings of the ACL SIGDAT-Workshop: 1995.
- Schmid, Helmut: Deep Learning-Based Morphological Taggers and Lemmatizers for Annotating Historical Texts. In: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage DaTeCH. Brussels, Belgium: 2019.

Software:

spacy , flair, TreeTagger, RNNTagger, SoMeWeTa

Projekte:

SoMeWeTa, Liste von Part of Speech Taggern

Verweise:

Part-of-Speech-Tagging, Tagsets, spaCy, WebLicht, Lemmatisierung, Tokenizer

Themen:

Natural Language Processing

Zitiervorschlag:

Eder, Elisabeth. 2021. Tagger. In: KONDE Weißbuch. Hrsg. v. Helmut W. Klug unter Mitarbeit von Selina Galka und Elisabeth Steiner im HRSM Projekt "Kompetenznetzwerk Digitale Edition". URL: https://gams.uni-graz.at/o:konde.176